The Canadian Geographer
Le Géographe canadien
Canadian Association of Geographers
Association Canadienne des Géographes

A new tool for neighbourhood change research: The Canadian Longitudinal Census Tract Database, 1971–2016

Jeff Allen
Department of Geography & Planning, University of Toronto

Zack Taylor
Department of Political Science, University of Western Ontario

Key Messages

- Neighbourhood change research is challenged by census boundaries being revised each census year.
- This paper describes the creation of a longitudinal spatial database of census tracts in Canada, bridging tract-level data for the 1971–2016 quinquennial censuses to a common set of boundaries.
- Methodology includes map-matching, dasymetric overlays, and population-weighted areal interpolation in order to minimize error when boundaries change over time.

Performing longitudinal analysis of socio-economic change in small-area spatial units such as census tracts presents several methodological complications and requires significant data preparation. Unit boundaries are revised each census year because of changes in population and delineation methodologies. This limits cross-year comparison since data are not representative of the same spatial units. To address these problems, we have developed an innovative procedure to reduce error when comparing tract-level data across census years by apportioning data to the same areal units. This paper describes the methods used to create the Canadian Longitudinal Tract Database. Our procedure is a combination of map-matching techniques, dasymetric overlays, and population-weighted areal interpolation. The output is a set of tables with apportionment weights pertaining to pairs of unique boundary identifiers across census years, which can be linked with census data or other data with census identifiers that require longitudinal comparison.

Keywords: census tracts, GIS, areal interpolation, neighbourhood change, historical geography

Un nouvel outil pour la recherche sur les changements dans le voisinage : la base de données longitudinales canadiennes des secteurs de recensement, 1971–2016

La réalisation d'une analyse longitudinale des changements socioéconomiques dans des unités spatiales de zone restreinte, par exemple des secteurs de recensement, présente plusieurs complications méthodologiques et requiert une préparation importante de données. Les limites des unités sont révisées chaque année de recensement en raison des changements dans la population et dans les méthodologies de délimitation. Ceci limite la comparaison des différentes années puisque les données ne sont pas représentatives des mêmes unités spatiales. Pour résoudre ces problèmes, nous avons développé une procédure innovatrice pour réduire les erreurs lors de la comparaison des données au niveau des secteurs entre les années de recensement en affectant les données aux mêmes unités spatiales. La présente communication décrit les méthodes utilisées pour créer la base de données longitudinales canadiennes des secteurs de recensement. Notre procédure est

Correspondence to / Adresse de correspondance: Jeff Allen, Department of Geography & Planning, University of Toronto, Sidney Smith Hall, 100 St. George Street, Toronto, ON M5S 3G3. Email/Courriel: jeff.allen@mail.utoronto.ca

The Canadian Geographer / Le Géographe canadien 2018, xx(xx): 1–14
DOI: 10.1111/cag.12467

une combinaison de techniques de corrélation de cartes, de superpositions dasymétriques et d'interpolation spatiale pondérée selon la population. Le résultat est une série de tableaux avec une répartition des pondérations se rapportant à des paires d'identificateurs uniques des limites pour chacune des années de recensement qui peuvent être reliées aux données du recensement ou à d'autres données avec des identificateurs du recensement qui requièrent une comparaison longitudinale.

Mots clés : secteurs de recensement, SIG, interpolation spatiale, changement dans le voisinage, géographie historique

Introduction

Canada's Census is a crucial research infrastructure used by scholars, governments, non-governmental organizations, and private-sector firms. Every five years, Statistics Canada collects a wide range of information about individual and household characteristics and behaviour from a representative sample of the national population. These data are disseminated as aggregate counts pertaining to multiple geographical units—provinces, subprovincial units known as census divisions and census subdivisions, metropolitan areas, federal electoral districts, and neighbourhood-sized census tracts—and as anonymized microdata. These data are more easily analyzed cross-sectionally rather than longitudinally because the boundaries of small-area spatial units such as census tracts are revised each census year to account for changing populations. At the same time, data formats, projections, and levels of precision have changed as technology has improved. Rectifying these inconsistencies is a laborious process that a number of scholars have undertaken on an ad hoc basis for specific projects, both in Canada (e.g., Schuurman et al. 2006; Walks and Maaranen 2008; Hulchanski 2010) and elsewhere (e.g., Vrieling and Melser 2013).

In the United States (US), these problems have been comprehensively addressed by three major projects: sociologist John Logan's open Longitudinal Tract Data Base (LTDB), the open National Historic GIS (NHGIS), and private data vendor GeoLytics' Neighborhood Change Data Base (NCDB), each of which use interpolation procedures to bridge US census data from 1970 through 2010 to a common set of tract boundaries (see GeoLytics 2007; Logan et al. 2014). These datasets have been used creatively by scholars in multiple disciplines to analyze neighbourhood change over time, focusing on urban sprawl, racial and income inequality and segregation, immigration patterns, and other topics (e.g., Logan 2013; Delmelle 2015).

Our project aims to create a similar dataset for Canadian census tracts—an open-access research tool that will enable low-cost longitudinal neighbourhood-scale research by academic, public-sector, non-profit, and private-sector researchers alike. We apply techniques similar to those used to create the LTDB to build tables bridging tract-level data for the quinquennial censuses from 1971 to 2016. Our methods involve apportioning tract-level data between years using a combination of map-matching techniques, dasymetric overlays, and population-weighted areal interpolation. The product is a set of apportionment tables that link census tract identifiers across years, enabling the allocation of data from multiple census years to a common set of boundaries for analysis. While similar to published census concordance or correspondence tables that associate boundary identifiers for some census years to the one immediately previous, our tables are more functional because they contain apportionment weights. For example, if a tract splits into two parts, the weight indicates the proportion of the source tract's count to be allocated to each part. Our procedure for generating these tables employs open-source tools and is readily applicable to other levels of census geography or non-census spatial units. While the primary objective of the project is to create an open tool to simplify neighbourhood change research, it also demonstrates innovative techniques for relating data aggregated to different spatial units.

We begin with a general discussion of how census tracts are defined, how they have changed over time, and the challenges that complicate longitudinal tract comparison. Then, we describe our procedure and evaluate how it minimizes error when transferring data between boundaries. We conclude the paper with an outline of potential applications and directions for future work.

How census tracts are defined and revised

Census Tracts (CTs) are spatial units delineated by Statistics Canada that are designed to contain between 2,500 and 8,000 people (Statistics Canada 2017a). Their outer boundaries typically correspond to major roads and highways, railway corridors, watercourses and water bodies, and administrative boundaries. Most census variables are disseminated at the CT level. They are convenient units for neighbourhood-level socio-economic analysis because of their small population size, compact shape, and consistency of definition across cities. Statistics Canada only defines CTs for land within larger urban centres—designated Census Metropolitan Areas (CMAs) and some Census Agglomerations (CAs). In 2016, CTs covered 48 CMAs and CAs across Canada, with an average population of 4,576. (By contrast, the coverage of the U.S. Census Bureau's tract program has been national since 2000, and tract-like small-area units outside metropolitan areas called block numbering areas were defined in 1990.)

The census tract concept originated in the US at the turn of the 20th century when social surveyors saw the need to collect and analyze data for urban neighbourhoods. While eight large cities were tracted in the 1910 US Census, definitions and dissemination of data for census tracts was only fully standardized in the 1940 iteration. Canada first published census tract-level data for several metropolitan areas in 1951 (including population counts for 1941). The first digital release of spatial data was in 1971. While no digital CT boundaries were released in 1976, boundaries and data have been released digitally on a quinquennial basis since 1981. For census years prior to 1971, there have been some piecemeal attempts to digitize CTs from paper maps (see Walks and Maaranen 2008; Brittnacher and Lesack 2013), but no comprehensive national dataset exists. Table 1 provides a summary of the number of CMAs and CTs, and the population and land area of tracted areas for each census year from 1971 to 2016. Three-quarters of Canadians live in tracted areas, a proportion that has increased as the country has become more urbanized and Statistics Canada has expanded census tract coverage.

As illustrated in Figure 1, boundaries are routinely adjusted to reflect change in population and alterations to physical urban form. These changes fall into three main types: splits, merges, and "many-to-many" changes. (Logan et al. 2014 refer to these as splits, consolidations, and complex changes.) Tracts are split when the population grows or when they are subdivided by new linear features such as major roads, highways, or rail lines. Merges occur when the population declines or the linear features that define tract boundaries are removed. Many-to-many changes typically occur in sparsely populated rural areas. New census tracts are also created when boundaries are extended into previously untracted areas or when Statistics Canada designates new CMAs or CAs. A further type of change, minor adjustments to the boundaries of the same underlying feature, sometimes occurs in unpopulated areas such as parks, shorelines, and industrial areas, or when linear features such as roads are realigned. As these usually do not

Table 1
Summary of census tracts by census year.

Year	Number of CMAs	Number of CTs	Population in CTs	National population	% of national population in CTs	Land area in CTs (km^2)
2016	48	5,721	26,183,052	35,151,728	74%	228,100
2011	48	5,452	24,444,283	33,476,688	73%	150,300
2006	48	5,076	22,748,198	31,612,897	72%	147,300
2001	46	4,798	20,997,692	30,007,094	70%	135,500
1996	43	4,223	19,592,684	28,846,761	68%	97,500
1991	39	4,068	17,918,831	27,296,859	66%	81,800
1986	37	3,776	16,149,197	25,309,331	64%	84,400
1981	36	3,302	14,680,165	24,343,181	60%	53,000
1971	30	2,449	12,849,165	21,568,545	60%	32,000

Note: Land area in CTs is calculated from cartographic boundary files. As the treatment of water bodies and shorelines varies from year to year, numbers have been rounded to the nearest hundred.

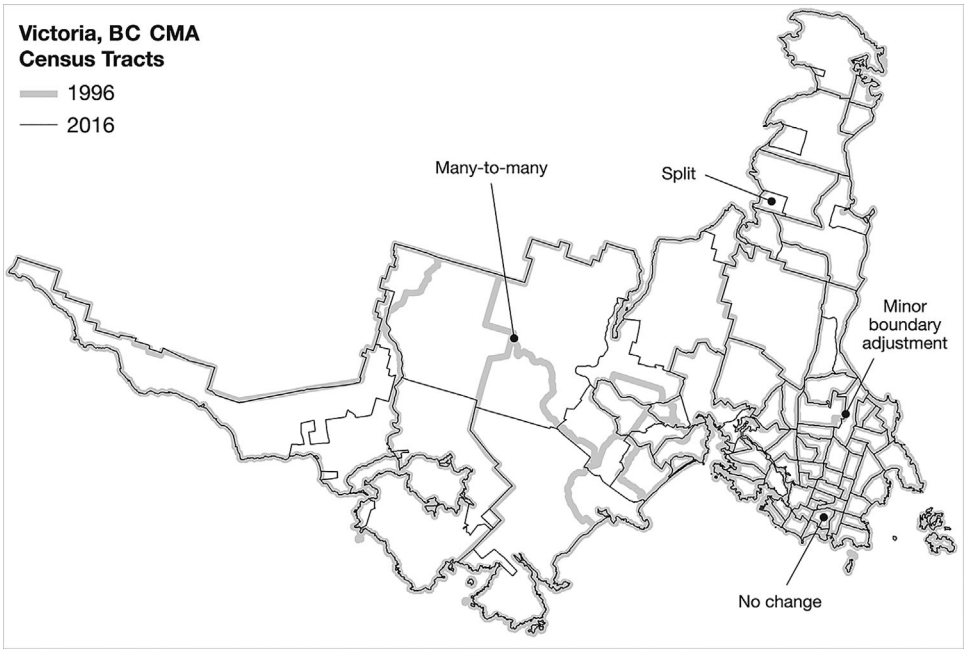

Figure 1
Boundary change in Victoria, B.C., 1996 and 2016. While most tract boundaries remained the same in the urban core, splits occurred in growing suburban neighbourhoods, and many-to-many changes are found in the sparsely populated wilderness areas west of the city.

affect the counts pertaining to each feature, these are treated as equivalent to "no change" in our analysis. Table 2 summarizes the prevalence of each type by census year pair. On average, for each new census year, the boundaries of 87% of census tracts remained unchanged, 4% split, and 9% were part of many-to-many changes. The number of splits and many-to-many changes is greater when comparing censuses that are further apart in time (e.g., there are more boundary changes comparing 1986 to 2016 than from 1986 to 1991).

Longitudinal comparison: Problems and solutions

The fundamental challenge when undertaking longitudinal comparison of spatial data is to ensure

Table 2
Summary of boundary changes by type and year.

	% of CTs			% of population in CTs			% of land area in CTs		
Year	**No change**	**Split**	**Many-to-many**	**No change**	**Split**	**Many-to-many**	**No change**	**Split**	**Many-to-many**
2011–2016	87%	3%	11%	84%	4%	12%	69%	10%	21%
2006–2011	88%	4%	7%	83%	9%	8%	66%	4%	29%
2001–2006	93%	3%	4%	90%	5%	4%	87%	3%	10%
1996–2001	84%	9%	7%	75%	17%	7%	72%	14%	15%
1991–1996	98%	0%	2%	98%	0%	2%	93%	2%	5%
1986–1991	81%	4%	16%	78%	7%	16%	55%	9%	36%
1981–1986	80%	6%	14%	75%	12%	14%	64%	8%	28%

that boundaries are consistent. When boundaries are intended to change over time—as is true of the census tract program—researchers need some way to relate data between years. An additional challenge faced by spatial analysts is that boundaries from different years often do not align, even when they have not meaningfully changed—that is, they represent the same underlying features, such as roads or watercourses. This is due to changes in the technical methodologies used to delineate boundaries of features between census years, including variation in projections, precision, and data formats (see Allen and Leahey 2016).

Areal interpolation

One common solution is *areal interpolation*: the process of transferring spatial data from one set of areal units (source zones, *s*) to another set of areal units (target zones, *t*). Different areal interpolation methodologies can be used to determine apportionment weights that indicate the proportion of data in a source zone that pertain to a given target zone. A common method for areal interpolation is *area weighting*, whereby an apportionment weight is calculated by dividing the area of the source zone that overlaps the target zone by the total area of the source (i.e., $w_{s,t} = a_{s \cap t} / a_s$). Goodchild and Lam (1980) originally used this method to link census tract data in London, Ontario to planning districts.

The problem with area weighting is that it assumes the uniform distribution of phenomena across the source zone, which is not usually the case when analyzing population. Accuracy of areal weighting of population data can be improved via *dasymetric mapping* techniques that categorize the source zones into classes before allocating portions to target zones (e.g., Eicher and Brewer 2001). Common dasymetric techniques include clipping out areas known to be unpopulated, such as water, greenspace, and industrial zones, and assigning data to specific land-use classes prior to computing weights.

Population weighting goes a step further. In this procedure, apportionment weights are generated using population counts pertaining to smaller spatial units. This has the same formulation as areal weighting but uses the intersection of the underlying population surface rather than area to generate weights: $w_{s,t} = p_{s \cap t} \ / \ p_s$. Population can be assigned to target zones either as points representing groups of people or tied to polygons combining population- and area-based weighting. Population weighting requires the availability of population counts for lower-level (smaller) units, which is not always the case. Canadian block-level population data are only available from 1991 onwards. One caveat about using population weighting is that it spuriously assumes that the spatial distribution of a sub-population group—for example, French-speakers or visible minorities—is identical to that of the overall population.

Figure 2 schematically demonstrates how these different interpolation strategies may produce different apportionment weights. Areal weighting assigns two-thirds of the source count to target tract 1 and one-third to target tract 2. Dasymetric areal weighting drops out unpopulated areas, of which there are more in tract 1. This produces equal weights for the two target tracts. Population weighting accounts for the density of the populated areas. As tract 2 has twice the population of tract 1, it is assigned a weight of two-thirds.

Partial existing solutions

There have been other attempts to link Canadian census data across years, however they are limited to single geographic areas and years of interest, vary in terms of sophistication, or have not been disseminated to enable replication, evaluation, and extension. The most open and comprehensive tool are Statistics Canada's correspondence tables, which link spatial unit identifiers only for adjacent census years (e.g., for 2006 to 2011, but not for 1991 to 2011). However, these tables do not include apportionment weights (i.e., they do not indicate the proportion of the source tract data that should be allocated to target tracts). Another strategy is to perform custom tabulations of census microdata to alternative spatial boundaries. This is rarely practicable, however, because access to microdata is typically limited to academic researchers, is subject to a rigorous application and vetting procedure, and sub-tract geographical locators are not always attached to individual records, especially in earlier census years. Canadian geographers have long been concerned about the relative inaccessibility of specialized data, including microdata, to non-profits and other civil society groups (Klinkenberg 2003).

Several academic neighbourhood change projects have independently employed various methods to

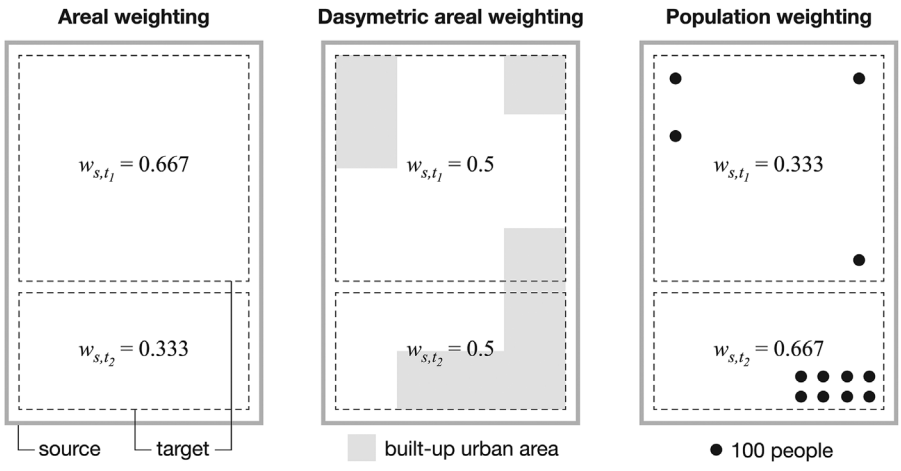

Figure 2
Schematic comparison of interpolation methods. The method used strongly influences the weights produced.

link census data across time. Murdie et al. (2014) and Simmons et al. (2009) have joined data associated with unique geographic unit identifiers across time, although the accuracy of this aspatial approach is questionable given the frequency of minor boundary changes and many-to-many relationships between units across years.

Another common method is to link data backwards in time by aggregating units. The most common type of boundary change in urban areas is the split. In this situation, count data from later years can simply be summed to the boundaries of earlier years' units. This strategy has been employed by recent reports on income polarization in several Canadian metropolitan areas from 1971 or 1981 to 2011 (Ley and Lynch 2012; Rose and Twigge-Molecey 2013; Prouse et al. 2014; Harris et al. 2015). It was also used by Ades et al. (2016), who aggregated 2006 data to 1986 boundaries to analyze changes in poverty distributions in Montreal, Toronto, and Vancouver. Again, this strategy breaks down in cases of minor boundary changes and many-to-many relationships. Also, the spatial specificity of the data may be diminished because aggregation across multiple census years yields large spatial units. This problem is especially acute in suburbanizing areas on the metropolitan fringe.

A third technique translates data from older source tracts to later target years through "inheritance." If unit boundaries are either split or remain unchanged over time, later units take on the value of the past unit to which they correspond. This approach has been used when analyzing non-count variables such as averages, medians, and percentages. Hulchanski (2010), for example, mapped neighbourhood income change in Toronto from 1970 to 2005 by apportioning all data to 2001 tract boundaries (see also Distatsio and Kaufman 2015). This approach is most effective in built-out urban areas like the city of Toronto, where the boundaries of the majority of tracts either remained consistent or split, minimizing complications associated with many-to-many changes. However, this process would be inadequate in areas featuring more complicated boundary changes.

Several projects have used the areal interpolation methods originally outlined by Goodchild and Lam (1980). For example, Walks and Maaranen (2008) used area-weighted interpolation to facilitate analysis of gentrification in Montreal, Toronto, and Vancouver between 1961 and 2001. Schuurman et al. (2006) conflated boundary mismatch and then used a population-weighting method to link socioeconomic data from 1996 to 2001 in the Vancouver region.

Importantly, each of these projects used idiosyncratic strategies to meet immediate research needs. Longitudinal neighbourhood change datasets were constructed containing specific variables, a select set of years, and focused on specific urban areas. In

most cases, the underlying datasets have not been made publicly accessible for use by other researchers. There is, however, precedent for the creation of open-access, national-scale datasets that disseminate census data for harmonized boundaries. The Canadian Century Research Infrastructure constructed historical spatial datasets for census divisions (i.e., counties and equivalents) and census subdivisions (i.e., municipalities) for census years spanning 1911 through 1951 (St-Hilaire et al. 2007). Despite evident interest in using census tract data longitudinally, ours is the first effort to build a comprehensive open-access tool for doing so.

American neighbourhood change databases

In the US, three major projects have used advanced areal interpolation methods to create longitudinal census tract spatial databases. The Longitudinal Tract Database (LTDB) used population-weighted areal interpolation at the block level for apportioning tract-level counts from 2000 to 2010, and areal interpolation of tract-level counts for 1970, 1980, and 1990, to create a series of apportionment tables that link data to 2010 boundaries (Logan et al. 2014). These tables were then used to create a series of synthetic data tables containing specific variable groups (e.g., income, ethnic diversity, education, etc.) that expedite neighbourhood change research.

Another project, the National Historical GIS (NHGIS), relates 1990 and 2000 blocks to 2010 blocks. The researchers developed a sophisticated dasymetric technique using road network and land cover data to generate apportionment weights (Schroeder 2017; see also McMaster et al. 2003; Schroeder 2007). Unlike the other products discussed here, the NHGIS does not use sub-tract units (blocks) to apportion counts pertaining to large units (tracts). In doing so, they can exploit the availability of block-level counts for variables other than total population (e.g., population counts by race and age, as well as household, family, and dwelling counts) to produce separate apportioned counts for those variables. As their entire process operates at the block level, they can aggregate counts to larger geographic units, including counties, tracts, and metropolitan areas.

The third is the Neighborhood Change Data Base (NCDB), a proprietary, for-profit application sold by GeoLytics, Inc., based on an earlier Urban Institute project funded by the Rockefeller Foundation (GeoLytics 2007). The GeoLytics application generates values for all comparable variables in the 1970, 1980, 1990, 2000, and 2010 census, apportioned to either 2000 or 2010 tract boundaries. Detailed information on its construction is not publicly available (but see Tatian 2003 for information on the earlier Urban Institute product). Logan et al. (2016) found that the 2000–2010 output was consistent with area-weighted interpretation with areas including both land and water areas.

Logan et al. (2016) analyzed the accuracy of these three American datasets by comparing their 2000 population counts in 2010 boundaries to U.S. Census Bureau tabulations. They found that deviance was greater in areas where boundaries changed over time, but that this could be minimized when using population-based rather than area-based weighting. (The LTDB and NHGIS were therefore found to be more accurate than the NCDB.) The NHGIS and LTDB projects directly inspired the creation of the Canadian Longitudinal Tract Database documented in this paper, with adjustments required due to differences in available Canadian census and ancillary data.

Methodology

The primary objective of this project was to generate a series of apportionment tables that can be used to transfer data across census years to a common set of boundaries, preserving as much fidelity as possible to observations' spatial location given available information. Apportionment tables are often called cross-reference or "crosswalk" tables in relational database management systems when examining many-to-many relationships. The apportionment tables generated for this project contain four fields: i_s, i_t, $w_{s,t}$, and $f_{s,t}$, where i_s is the unique identifier of the source CT and i_t is the unique identifier of the target CT. As these are Statistics Canada's standard identifiers, the apportionment tables are readily joined to other tabular data that use them. They can also be joined to spatial boundaries in a GIS to cartographically represent and spatially analyze linked data. The field $w_{s,t}$ is the associated weight, which can be used to apportion data from i_s to i_t. Weight values range from 0 to 1. Weights pertaining to a given source tract sum to 1. If $w_{s,t} = 1$ then the entirety of the data in the source tract, *s,* is apportioned into the target tract, *t.*

A dummy value of −1 is shown in cases where there is no source due to the creation of a new tract in a previously untracted area, or no target due to a reduction in tracted territory. Finally, the $f_{s,t}$ field contains flags indicating the type of boundary change (no change, split, many-to-many).

Since census data pertain to areal units, we employed methods of areal interpolation to generate the set of weights. The specific interpolation methods varied for each year depending on the format and availability of datasets. The tabular and spatial data used in this project originally came from Statistics Canada (1971 to 2016). Some of the census boundaries used came from a census boundary conversion project in which pre-2006 spatial datasets were converted from archaic data formats (e.g. e00) and flat text files into modern GIS formats such as ESRI Shapefiles (Allen and Leahey 2016). For the dasymetric part of our procedure, we also made use of an historical built-up area dataset produced by Statistics Canada. This dataset is provided as a 30m resolution grid and covers the majority of urban areas in Canada on a decennial basis back to 1971 (see Soulard et al. 2016 for a description of this dataset). The computational work for this project was conducted using custom functions written in the scripting language Python and using spatial queries in PostGIS, which is a geographic extension to the relational database management system PostgreSQL.

The remainder of this section details, in reverse chronological order, how the apportionment tables were generated for specific quinquennial census years spanning 1971 and 2016. The 1976 census is not included because Statistics Canada has not disseminated digital boundary files for that year.

Census years 2001, 2006, 2011, and 2016

Apportionment tables for the census boundaries in 2001, 2006, 2011, and 2016 were generated by combining dasymetric areal weighting with population-weighting interpolation using census dissemination blocks. Blocks are the smallest available subunits of all larger census units, including CTs. The first step was to remove parts of blocks assumed not to have any dwellings by clipping blocks using a water layer and intersecting them with the built-up area dataset. The second step was to compute a ratio of the block population, p_k, to the total population of the source census tract, p_s, to which it belongs. The third step was to spatially intersect the clipped blocks, k, with the target census tracts, t, and then compute a ratio of the area of intersection of the block by each target tract it overlays, by the total area of the block, $a_{k\cap t} / a_k$. Weights were created by multiplying these two ratios and then grouping the blocks upon rows with the same boundary identifiers in both years, summing the combined ratios—i.e., $w_{s,t} = \Sigma_{k\in K} [(a_{k\cap t} / a_k) (p_k / p_s)]$ where K is the set of blocks, k, that intersect the source tract, s, and the target tract, t. Note that Statistics Canada only releases population and dwelling counts at the block level. As a result, we cannot produce separate weights for other block-level sub-population counts, as in the NHGIS procedure.

Where possible, the blocks were intersected with Digital Boundary Files (DBF) rather than Cartographic Boundary Files (CBF). In most locations, CBFs are clipped to water features to provide more accurate cartographic representation. DBFs, by contrast, are typically not clipped to water features, which makes it less likely that portions of blocks will fall outside of desired unit boundaries.

Census years 1991 and 1996

The procedures for apportioning data for the census years 1991 and 1996 were complicated by incomplete coverage of block-level data in urban areas and extensive spatial mismatch between unit boundaries disseminated for these two census years compared to other census years' boundaries.

Block-level data for 1991 and 1996 are available as block-face points. Each block-face point represents one side of a road segment between two intersection nodes (e.g., a rectangular block would have four block-face points, one for each side). Their coordinates are typically offset by 10 to 20 metres from the street and include an address range, population count, and the unique identifiers of the higher-level census units with which it is associated. Importantly, block-face points are not available for all tracted areas. (Block-level data only became available for all of Canada in 2001.) In 1991 and 1996, 93% and 90% of census tracts, respectively, had block-face point coverage. Visual inspection indicates that tracts without block coverage are typically located in more rural areas. For tracted areas without block-face points, we generated a set of pseudo block-face points based on the location of

street segments. These were generated first by computing 100-metre buffers from street network files pertaining to 1991 and 1996, respectively; intersecting these buffers with the boundaries of Enumeration Areas (EAs are smaller areas than CTs but larger than the blocks produced in later years); and then generating n random points within these intersected areas, where n is the population of the EA. The assumption that dwellings are more likely to be located near roadways has been used in other areal interpolation projects (e.g., Reibel and Bufalino 2005). These ersatz block-face points were then merged with the existing block-face data.

The second major issue is the widespread spatial mismatch between 1991 and 1996 census boundaries and those of other years. This problem, which manifests as slippage and misalignment of feature boundaries, including roads, rivers, and shorelines, is the product of varying digitization procedures, precision, and projections used for delineating census boundaries in different parts of the country. The problem is not easily fixed because the character and magnitude of the mismatch varies between, and even within, CMAs. In some parts of the country, it is inconsequential, whereas in others the boundaries are offset by more than 200 metres. If left uncorrected, spatial mismatch would result in extensive error when conducting the spatial intersection stage of the apportionment process. Many block-face points located near boundary edges would be incorrectly allocated to adjacent units. Others have recognized and proposed solutions to this problem. For example, Schuurman et al. (2006) conflated 2001 geometry to the offset 1996 boundaries in Vancouver in order to link data from 1996 to 2001.

We resolved this problem with our own conflation procedure that uses a spatially weighted average to translate each coordinate in the 1991 and 1996 census spatial files in relation to a much smaller set of manually created control points. The first step was to generate control points, each of which contains the coordinates of a stable known point—a street intersection—according to the original 1996 or 1991 census boundaries, and the coordinates of where it should be located according to census boundaries of later years. Each control point therefore quantifies the amount of spatial mismatch for a particular location. The more control points, the greater the accuracy of the procedure. Generating and checking the control points was partly a manual process, so this was limited to roughly 2,000 points for all census tract coverage in Canada. More control points were generated in urban areas where there was greater mismatch.

After generating the control points, the coordinates in the boundary files were translated using a distance-weighted average of the mismatch held in nearby control points. Updated coordinates were generated via the formulas $x_j = \Sigma_c\,(q_{ic}\,\Delta x_c) + x_i$ and $y_j = \Sigma_c\,(q_{ic}\,\Delta y_c) + y_i$, where q_{ic} is the distance-based weight for each control point and Δx_c and Δy_c are the mismatch in the control points. The distance-based weighted q_{ic} can be found via $q_{ic} = {d_{ic}}^{\beta} / \Sigma_c\, {\mathrm{d}_{\mathrm{ic}}}^{\beta}$ where d_{ic} is the straight-line distance between the coordinates of an input point, i, and the coordinates of the control point, c, and β is a decay parameter (see Figure 3 for a schematic of the procedure). The value of β was estimated to be -2 after systematically visually inspecting numerous values. These functions were built in a custom Python script and set up to loop over each input point in the sets of 1991 and 1996 boundaries and block-face points across Canada, grouping by year and CMA. In order to preserve topology, tract boundaries were converted to the open TopoJSON format prior to running the conflation procedure. In this format, each boundary is stored with a set of unique node identifiers pertaining to a list of coordinates. As adjacent boundaries share node identifiers, coordinates are only translated once, even when they pertain to more than one boundary polygon. Once conflated, a population-weighted interpolation procedure was used to generate apportionment tables linking with other census years. The first step of this procedure was to compute a ratio of the block-face population, p_k, to the total population of the census tract, p_s, in which it belongs. The second step was to spatially join the unique identifiers of the target census tracts, t, to the set of translated block points, k. Weights were created by grouping the block points upon rows with the same boundary identifiers in both years, summing their ratios—i.e., summing every ratio p_k / p_s where k intersects both s and t.

Census years 1981 and 1986

Block-level population counts were not publicly disseminated for the 1981 and 1986 censuses, nor was there national coverage of enumeration areas. Without small-area population data, we used dasymetric areal interpolation to generate the apportionment weights for these years. This involved

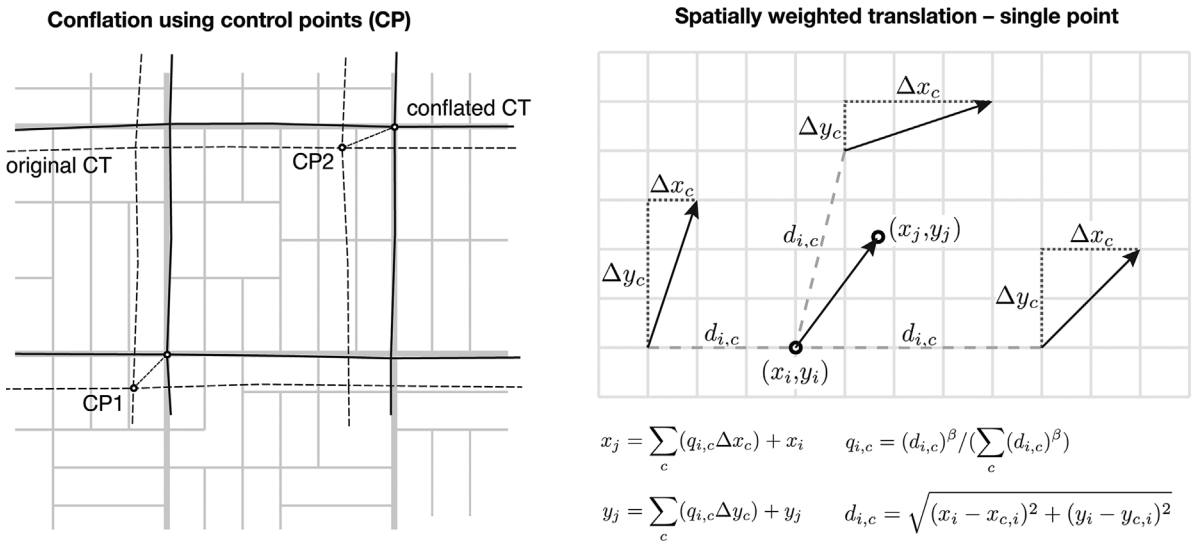

Figure 3
Schematic diagram of the conflation procedure. First, control points referencing known stable locations were identified in the 1991 and 1996 boundary files (left). Then, the magnitude of the translation between them was calculated using a process of distance-based weighting (right).

clipping out areas in census tracts assumed to have no dwelling units via a water layer and the built-up area dataset. Once clipped, areal weights were determined by computing the ratio of the intersecting area to the total area of the source tract. Minor misalignment of boundaries across years generated thousands of slivers along tract edges. Most occurrences are spurious and can be removed. Others are the product of minor boundary changes, such as road realignments. A sliver was removed if its weight was less than 0.05 and the occurrences of its source and target CTs were greater than one. The value of 0.05 was determined after testing multiple values and seeing which minimized error. The procedure for estimating error is explained in the following section. Removing the slivers results in the sum of weights for a source tract equalling less than 1. To resolve this, the differences were calculated ($\delta_s = 1 - \Sigma_s\ w_{s,t}$) and then equally distributed among the remaining weights ($w_{s,t} = w_{s,t} + \delta_s\ /\ n_s$ where n_s is the number of weights in the apportionment table pertaining to the source tract, s).

Census year 1971

Unlike 1981 and 1986, lower-level spatial boundaries are available for 1971 via Statistics Canada's recently restored EA boundary file (Statistics Canada 2017b). These polygons cover the majority of urban regions in 1971 and were digitized in reference to 2011 street network files. Because of population growth since 1971, several of the EA boundaries are as large as, or larger than, CTs in later years. A combined population and area-weighting approach was used to interpolate data from 1971 to later years. For each 1971 EA, we computed the ratio of the EA population to the total population of its associated 1971 CT. The EA polygons were then intersected with the geometry of the target tracts, *t*. A second ratio was then calculated of the intersection area to the total area of the EA. The weights were generated by multiplying these area and population ratios, and then summing them, grouping by the pairs of source tracts and target tracts.

Generating flags indicating join type

The apportionment tables include flags ($f_{s,t}$) indicating the type of spatial join connecting each source and target tract. These flags also allow for an overall comparison of how census boundaries have changed between census years. The flags were generated using a custom Python script that loops over an apportionment table and counts the

instances of each unique source identifier and each target identifier. It then assigns flags to each source-target row based on their corresponding counts. If a source tract appears only once in the apportionment table, and its associated target zone only appears once, it is flagged as a "one-to-one" association (i.e., no boundary change has occurred, $f_{s,t}=1$). If multiple source tracts correspond to a single target tract, they are flagged as a "merge" ($f_{s,t}=2$). If a single source tract corresponds to multiple target tracts, each with a count of 1, it is flagged as a "split" ($f_{s,t}=3$). All others are flagged as "many-to-many" relationships ($f_{s,t}=4$).

Validating results

We validated our procedure by comparing population counts apportioned from our interpolation process with Statistics Canada's adjusted counts. Adjusted counts are found in a published dataset of the population counts of the previous census year estimated to the current year's boundaries. The difference between our interpolation procedure and these adjusted counts is not a true error measure, as the adjusted counts potentially include their own uncertainties due to Statistics Canada's geocoding methodology and spatial aggregation procedures. Nevertheless, they are a viable means of validating our estimates.

We computed frequencies of relative deviance and root-mean-square deviation (RMSD). Relative deviance, or percent deviance, is the ratio of the absolute difference of the estimated and reference value, normalized by the reference value, for each census tract. The RMSD is a measure of the overall deviance for the entire set of tracts that are being apportioned. $\text{RMSD} = [(y_i^* - y_i)^2 / n]^{0.5}$ where y_i^* is the adjusted count for a tract, y_i is the apportioned count from our procedure, and n is the total number of census tracts in the target year. The RMSD counts large absolute differences disproportionately, compared to small differences, since they are squared before being summed. These measures of deviation for each pair of years back to 1981 are summarized in Table 3, showing how the interpolation method we employed compares with basic areal weighting. As Statistics Canada's adjusted counts are only available for adjacent year pairs, error in our methodology can only be

Table 3
Deviance estimates by year and interpolation method.

	Frequencies of relative error						
Methodology by year-pair	**< 0.01%**	**0.01–0.99%**	**1.00–2.99%**	**3.00–4.99%**	**5.00–9.99%**	**≥ 10.00%**	**RMSD**
2011 to 2016							
Population weighting	0.909	0.051	0.019	0.007	0.007	0.007	44
Areal weighting	0.319	0.396	0.148	0.039	0.029	0.070	735
2006 to 2011							
Population weighting	0.886	0.055	0.025	0.013	0.011	0.011	61
Areal weighting	0.390	0.340	0.105	0.030	0.024	0.111	898
2001 to 2006							
Population weighting	0.955	0.029	0.007	0.002	0.003	0.004	36
Areal weighting	0.710	0.182	0.019	0.008	0.012	0.070	783
1996 to 2001							
Population weighting	0.779	0.059	0.051	0.030	0.032	0.048	183
Areal weighting	0.005	0.204	0.272	0.135	0.137	0.246	1298
1991 to 1996							
Population weighting	0.958	0.014	0.005	0.004	0.005	0.014	107
Areal weighting	0.217	0.708	0.031	0.013	0.010	0.020	334
1986 to 1991							
Dasymetric areal weighting	0.748	0.013	0.028	0.024	0.042	0.146	586
Areal weighting	0.003	0.210	0.289	0.159	0.153	0.187	730
1981 to 1986							
Dasymetric areal weighting	0.705	0.022	0.022	0.023	0.041	0.187	853
Areal weighting	0.003	0.162	0.281	0.146	0.150	0.258	1234

estimated on this basis (i.e., from 1981 to 1986, not from 1981 to 2016). It is expected that interpolating over a greater time period would result in more error as there are more boundary changes. Table 4 exemplifies how the frequency of deviations varies in the 2011–2016 year-pair between boundary change types for multiple interpolation procedures. This table shows that deviance is greater when boundaries change, but is reduced when combining population and dasymetric procedures. The measures computed in Tables 3 and 4 are of similar magnitude to those calculated by Logan et al. (2016) for American census tracts.

Output and use

We have generated apportionment tables for all pairs of adjacent census years (e.g., 1986 to 1991) and for each year to 2016 (e.g., 1986 to 2016) to link all data to this common set of boundaries. Our apportionment tables are provided as comma-separated value (.csv) tables so they can be imported into nearly any open-source or proprietary GIS, database management system, or spreadsheet software. The tables and supporting materials are available on the Scholars Portal Dataverse (https://doi.org/10.5683/SP/EUG3DT). Sample code for apportioning data is posted on GitHub (https://github.com/jamaps/CLTD).

Count variables (e.g., population) are simplest to work with. They can be apportioned by joining the data on the source tract identifiers, multiplying the variable by the weights, and then grouping by the target tract identifiers, summing any apportioned data: $y_t = \Sigma_s\, y_s\, w_{s,t}$.

Non–count variables that represent means, ratios, or percentages (e.g., average household income, percentage below low-income cut-off, or unemployment rate) require additional steps because they are abstracted from the underlying counts. Where possible, our solution is to convert them into

Table 4
Comparison of deviance by boundary change type and interpolation method, 2011–2016.

	Frequencies of relative error						
Change type	**< 0.01%**	**0.01–0.99%**	**1.00–2.99%**	**3.00–4.99%**	**5.00–9.99%**	**≥ 10.00%**	**RMSD**
Population + dasymetric weighting							
No change	0.988	0.008	0.003	0.000	0.000	0.000	14
Split	0.511	0.314	0.057	0.030	0.030	0.057	76
Many-to-many	0.512	0.252	0.120	0.045	0.045	0.027	116
All	**0.909**	**0.051**	**0.019**	**0.007**	**0.007**	**0.007**	**44**
Population + areal weighting							
No change	0.824	0.036	0.103	0.022	0.011	0.004	49
Split	0.489	0.339	0.093	0.032	0.025	0.022	125
Many-to-many	0.366	0.177	0.220	0.100	0.072	0.065	223
All	**0.757**	**0.067**	**0.116**	**0.031**	**0.018**	**0.012**	**90**
Population block centroid weighting							
No change	0.997	0.002	0.001	0.000	0.000	0.000	2
Split	0.905	0.020	0.014	0.014	0.007	0.041	191
Many-to-many	0.570	0.160	0.139	0.052	0.047	0.032	163
All	**0.944**	**0.021**	**0.017**	**0.007**	**0.006**	**0.006**	**70**
Dasymetric areal weighting							
No change	0.988	0.005	0.003	0.001	0.001	0.002	71
Split	0.007	0.022	0.063	0.042	0.127	0.739	1855
Many-to-many	0.755	0.053	0.027	0.021	0.027	0.118	573
All	**0.913**	**0.011**	**0.009**	**0.005**	**0.010**	**0.052**	**463**
Areal weighting							
No change	0.382	0.425	0.132	0.029	0.016	0.016	398
Split	0.000	0.027	0.074	0.044	0.091	0.764	2503
Many-to-many	0.005	0.343	0.297	0.110	0.097	0.148	851
All	**0.319**	**0.396**	**0.148**	**0.039**	**0.029**	**0.070**	**735**

counts by multiplying the percentages and denominators by the appropriate denominator, as follows: $r_t = (\Sigma_s\ r_s\ y_s\ w_{s,t})\ /\ (\Sigma_s\ y_s\ w_{s,t})$, where r_s is the variable pertaining to the source census tract, y_s is the population of the group to which this variable refers, and r_t is the value in the target CT. For example, if r_s is average household income, y_s would be the total number of households. Variables representing medians cannot be apportioned accurately when boundaries change because the underlying distribution of observations is not known.

Users of the tables should be mindful that tract relationships only exist where tracts exist in the source year. The 1971-to-2016 table, for example, will only apportion data for territory that was tracted in 1971. As noted in Table 1, tract coverage has increased almost fivefold between 1971 and 2016. Of this increase, 58% occurred through the expansion of the metropolitan areas designated in 1971; the remaining 42% occurring through the initial tracting and later expansion of additional urban agglomerations. Depending on the researcher's objective, this may or may not be relevant. We recommend exploratory mapping to ensure that the phenomenon of interest falls within the area covered by the apportionment tables.

Conclusion and future work

Canadian Census products are not designed for convenient longitudinal analysis. Using them in this way requires significant effort to transform the data. To date, researchers have used a variety of methods to do this. We created the Canadian Longitudinal Tract Database to accomplish three goals. First, we hope the tables and scripts' open availability and ease of use will unlock new possibilities for academic, professional, and community-based researchers interested in analyzing neighbourhood change. Second, we hope it will become a standard tool and, in so doing, make the outputs of different projects directly comparable. All of the tools and libraries used are open-source, and all custom scripts created for this project are made available on GitHub (https://github.com/jamaps/CLTD). Using open-source software and publicly disseminating our code allows for our work to be reproduced and improved upon by others. We ask only that errors and improvements be reported back to the authors.

Third, we sought to develop, drawing on best international practices, an innovative set of methods for spatial apportionment that can be generally applied under a range of conditions, including constraints on source data availability. As we have demonstrated, error is minimized by combining population-based, areal, and dasymetric procedures. These procedures are readily applied to other census boundaries—such as forward sortation areas, census subdivisions, dissemination areas, the newly introduced aggregate dissemination areas, or dissemination blocks—to permit longitudinal analysis at other geographic scales. The territorial coverage of our work could also be extended if Statistics Canada were to disseminate internal spatial datasets, including EA boundaries for 1981 and 1986, or boundaries and associated data for the provincial census tracts (PCTs) defined for all territory outside of census metropolitan areas between 1971 and 1991.

We foresee two directions for future work. The first is to develop more refined interpolation techniques to minimize error. This could include improving the translation procedure for 1991 and 1996 boundaries, or inputting more accurate land-use classifications for dasymetric interpolation. In addition, we could explore the creation of alternative weights for apportioning non-population variables—households, families, and dwellings—for census years where lower-level units and counts are available.

A second avenue would be to further develop the deviance estimates. Uncertainty increases when boundaries change, so error is most likely magnified as the time period being bridged becomes longer. Directly measuring this is not currently possible, however, because our procedure validates results in relation to Statistics Canada's adjusted counts, which are only available for adjacent census years. A possible remedy would be to purchase custom tabulations of pre-2011 population data in 2016 boundaries. Another extension would be to examine potential error of areal interpolation of specific variables, not just population, as the spatial distribution of every variable differs. Much like the American LTDB and GeoLytics products, our tables assume that subpopulation groups have the same spatial distribution as the overall population.

Acknowledgements

This project is supported by an Insight Development Grant from the Social Sciences and Humanities Research Council of Canada, file no. 430-2016-00331, and by a Faculty Research Development Grant from the Dean of Social Sciences, University of Western Ontario.

References

Ades, J., P. Apparicio, and A. Séguin. 2016. Is poverty concentration expanding to the suburbs? Analyzing the intra-metropolitan poverty distribution and its change in Montreal, Toronto and Vancouver. *Canadian Journal of Regional Science* 38(1/3): 23–37.

Allen, J., and A. Leahey. 2016. Improving access to digital historical census boundaries in Canada. *ACMLA Bulletin* 153: 44–50.

Brittnacher, T., and P. Lesack. 2013. *Census of Canada. Boundary files, 1951 [2013].* Vancouver, BC: University of British Columbia Library. Humanities and Social Sciences Division. Data Services. http://hdl.handle.net/11272/10268.

Delmelle, E. C. 2015. Five decades of neighborhood classifications and their transitions: A comparison of four US cities, 1970–2010. *Applied Geography* 57: 1–11.

Distatsio, J., and A. Kaufman, eds. 2015. *The divided prairie city: Income inequality among Winnipeg's neighbourhoods, 1970–2010.* Winnipeg, MB: Institute of Urban Studies, University of Winnipeg.

Eicher, C., and C. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28(2): 125–138.

GeoLytics, Inc. 2007. *CensusCD neighborhood change database (NCDB) 1970–2000 US Census Tract Data.* East Brunswick, NJ: GeoLytics, Inc.

Goodchild, M., and N. Lam. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-processing* 1: 297–312.

Harris, R., J. Dunn, and S. Wakefield. 2015. *A city on the cusp: Neighbourhood change in Hamilton since 1970.* Toronto, ON: Cities Centre, University of Toronto.

Hulchanski, J. D. 2010. *The three cities within Toronto: Income polarization among Toronto's neighbourhoods, 1970–2005.* Toronto, ON: Cities Centre, University of Toronto.

Klinkenberg, B. 2003. The true cost of spatial data in Canada. *Canadian Geographer* 47(1): 37–49.

Ley, D., and N. Lynch. 2012. *Divisions and disparities in lotus-land: Socio-spatial income polarization in greater Vancouver, 1970–2005.* Toronto, ON: Cities Centre, University of Toronto.

Logan, J. 2013. The persistence of segregation in the 21st century metropolis. *City & Community* 12(2): 160–168.

Logan, J., B. Stults, and Z. Xu. 2016. Validating population estimates for harmonized census tract data, 2000–2010. *Annals of the American Association of Geographers* 106(5): 1013–1029.

Logan, J., Z. Xu, and B. Stults. 2014. Interpolating U.S. decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer* 66(3): 412–420.

McMaster, R., M. Lindberg, and D. Van Riper. 2003. The national historical geographic information system (NHGIS). In *Proceedings, 21st International Cartographic Conference.* International Cartographic Association, 821–828.

Murdie, R., R. Maaranen, and J. Logan. 2014. *Eight Canadian metropolitan areas: Spatial patterns of neighbourhood change, 1981–2006: A typology based on a combined statistical analysis of census tract data.* Toronto, ON: Cities Centre, University of Toronto.

Prouse, V., J. Grant, M. Radice, H. Ramos, and P. Shakotko. 2014. *Neighbourhood change in Halifax Regional Municipality, 1970 to 2010: Applying the "three cities" model.* Toronto, ON: Cities Centre, University of Toronto.

Reibel, M., and M. E. Bufalino. 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 37(1): 127–139.

Rose, D., and A. Twigge-Molecey. 2013. *A city-region growing apart? Taking stock of income disparity in greater Montréal, 1970–2005.* Toronto, ON: Cities Centre, University of Toronto.

Schroeder, J. 2007. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39(3): 311–335.

——. 2017. Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. *Computers, Environment and Urban Systems* 62: 53–63.

Schuurman, N., D. Grund, M. Hayes, and S. Dragicevic. 2006. Spatial/temporal mismatch: a conflation protocol for Canada census spatial files. *Canadian Geographer* 50(1): 74–84.

Simmons, J., L. Bourne, and S. Kamikihara. 2009. *Changing economy of urban neighbourhoods: An exploration of place of work data for the greater Toronto region.* Toronto, ON: Cities Centre, University of Toronto.

Soulard, F., G. Gagnon, and J. Wang. 2016. *The changing landscape of Canadian metropolitan areas, human activity and the environment.* Ottawa, ON: Statistics Canada.

St-Hilaire, M., B. Moldofsky, L. Richard, and M. Beaudry. 2007. Geocoding and mapping historical census data: The geographical component of the Canadian century research infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 40(2): 76–91.

Statistics Canada. 2017a. *Dictionary, Census of Population, 2016.* http://www12.statcan.gc.ca/census-recensement/2016/ref/dict/index-eng.cfm.

——. 2017b. *Restoration of the 1971 enumeration area polygons for Canada's largest cities.* http://www.statcan.gc.ca/pub/16-510-x/16-510-x2017001-eng.htm.

Tatian, P. A. 2003. *Neighborhood change database–NCDB. 1970–2000 tract data: Data users guide.* Washington, DC: Urban Institute.

Vrieling, A., and C. Melser. 2013. Constructing boundary-consistent population time series for the municipalities of the Netherlands, 1988–2011. *Population Studies* 67(2): 195–208.

Walks, R. A., and R. Maaranen. 2008. *The timing, patterning and forms of gentrification and neighbourhood upgrading in Montreal, Toronto, and Vancouver 1961 to 2001.* Toronto, ON: Cities Centre, University of Toronto.